CAPÍTULO 2

Análisis de varianza



INTRODUCCIÓN

El análisis de varianza (*analysis of variance*, Anova) es una herramienta estadística que permite probar la igualdad de tres o más medias poblacionales con los datos obtenidos con muestras de cada una de estas poblaciones.

Con el análisis de varianza podemos hacer inferencias acerca de si nuestras muestras se tomaron de poblaciones que tienen la misma media. Algunos ejemplos donde podemos aplicar este análisis son:

- En una universidad, deseamos comparar cuál de cuatro métodos distintos de enseñanza-aprendizaje de estadística (presencial, autodidacta, multimedia, sistema abierto) permite obtener un aprendizaje más rápido de la materia.
- Para una empresa que produce tres tipos de gasolina sin azufre (Nova, Magna y Premium) comparar el rendimiento del kilometraje producido en motores de combustión interna.
- Comparar los ingresos recibidos el primer año por los egresados de cuatro universidades públicas de la Ciudad de México (unam, uam, ipn, ucm).
- Medir si la concentración de contaminantes en la ciudad de Monterrey presenta características similares o distintas en cada 1 de las 4 estaciones del año, lo cual permitirá determinar políticas públicas para el control de contaminantes en la ciudad.

El análisis de cada factor podría hacerse por pares, por ejemplo, en el caso del rendimiento del kilometraje por tipo de gasolina, las comparaciones podrían hacerse como: Nova-Magna, Nova-Premium y Magna-Premium. Lo que representa desarrollar pruebas de hipótesis con dos muestras,[1] pero aquí, la prueba estadística comprende como ya indicamos tres o más muestras a la vez (Nova-Magna-Premium).

Como puede observarse de los ejemplos, el Anova consistiría en hacer pruebas para encontrar las diferencias entre las medias poblacionales. Para el ejemplo de los rendimientos de kilometraje obtenido con las diversas gasolinas, comparamos el rendimiento de la gasolina Nova con el rendimiento de la gasolina Magna y con el rendimiento de la gasolina Premium; este análisis implica un examen de las varianzas (o variaciones) de cada muestra, de aquí el nombre del procedimiento estadístico de análisis de varianza. La inferencia obtenida nos indicará si los rendimientos de cada combustible son iguales o diferentes.

Otra de las aplicaciones del análisis de varianza es el análisis de los resultados en un análisis de regresión en donde por un lado se cuenta con datos experimentales y por otro con datos producto de una observación al mismo tiempo.

Conceptos básicos

El principal objetivo del análisis de varianza es identificar el factor o los factores que producen la variabilidad en un conjunto de datos.

Si un solo factor (variable independiente) es el que produce esta variabilidad, entonces el análisis de varianza recibe el nombre de *análisis de varianza de un solo factor* o *One-way* Anova.

Cada factor está asociado con un conjunto particular de datos (muestra) o tratamientos (variable dependiente). El término tratamiento se origina de la agricultura, ya que en ésta se utilizó por primera vez este procedimiento estadístico. Los tratamientos se derivaron de tratar varias parcelas de tierra con diferentes fertilizantes para medir el factor de rendimiento promedio por cultivo.

Suponga ahora que obtenemos una muestra aleatoria para el tratamiento i $[x_{i1}, x_{i2}, x_{i3}, ..., x_{ij}]$. Entonces, $i = 1, 2, 3, \ldots, n$ se refiere a cada tratamiento o muestra en el problema y $j = 1, 2, 3, \ldots, n$, a cada dato que forma la muestra o tratamiento (*véase* el cuadro 2.1).

[1] Véase el tema de “Pruebas con dos muestras” en Rodríguez, J. Pierdant, A. y Rodríguez, C. (2008), *Estadística para administración*, Grupo Editorial Patria, México, pp. 405-420.

Cuadro 2.1 Cuadro estadístico que muestra los tratamientos y sus datos.

	Datos del tratamiento				
Tratamiento	1	2	3	…	j
1	X_{11}	X_{12}	X_{13}	…	X_{ln}
2	X_{21}	X_{22}	X_{23}	…	X_{2n}
3	X_{31}	X_{32}	X_{33}	…	X_{3n}
4	X_{41}	X_{42}	X_{43}	…	X_{4n}
…	…	…	…	…	X_{ln}
…	…	…	…	…	…
I	X_{I1}	X_{I2}	X_{I3}	…	X_{In}

Es necesario observar que en este caso cada tratamiento i tiene el mismo número de datos n, pero el modelo se puede generalizar cuando existen tratamientos que presentan diferente número de datos.

Por tanto, en este procedimiento estadístico se requieren tres supuestos:

- Cada tratamiento (población) está normalmente distribuida.
- Todos los tratamientos (poblaciones) tienen la misma varianza $[(\sigma)^2]$.
- Las muestras de cada tratamiento se seleccionan independientemente.

Con base en estos supuestos, el análisis de varianza consiste en decidir si las "i" muestras o tratamientos se tomaron de poblaciones que tienen la misma media μ. Para ello se establecen dos hipótesis de prueba.[2]

- H_0: $\mu_1 = \mu_2 = ... \mu_I$, es decir, la media de todas las poblaciones es igual.
- H_1: $\mu_1 \neq \mu_2 \neq ... \mu_I$, es decir, las medias poblacionales son distintas.

El análisis de varianza, por tanto, nos indicará si las medias de todas las poblaciones son iguales; es decir, no hay una variación significativa entre ellas, o bien, si entre éstas hay una variación significativa, ya que este comportamiento puede presentar alguna de las cuatro características siguientes.

1. No hay variación entre poblaciones (tratamientos) ni dentro de cada una de ellas.

Consideremos el ejemplo de las gasolinas, probando su rendimiento (kilómetros por cada litro) en doce motores de combustión interna y se obtienen los resultados que se muestran en el cuadro 2.2.

Cuadro 2.2 Probando el rendimiento de las gasolinas en kilómetros por litro.

	Kilómetros por litro				
Tratamiento	Motor 1	Motor 2	Motor 3	Motor 4	Promedio
Nova	10	10	10	10	10
Magna	10	10	10	10	10
Premium	10	10	10	10	10
			Promedio de los promedios =		10

2. Hay variación entre poblaciones (tratamientos), pero no dentro de cada una de ellas.

Continuando con el ejemplo del rendimiento por kilómetro de cada gasolina, los resultados para este caso se muestran en el cuadro 2.3.

[2] Recuerde que H_0 representa la hipótesis nula y H_1 la hipótesis alternativa. Le sugerimos revisar el tema de Pruebas de Hipótesis en Rodríguez J., Pierdant A. y Rodríguez C. [2008], *Estadística para administración*, GE Patria, México, pp. 385-420.

Cuadro 2.3 Rendimiento por kiómetro de cada gasolina.

Kilómetros por litro					
Tratamiento	Motor 1	Motor 2	Motor 3	Motor 4	Promedio
Nova	9.5	9.5	9.5	9.5	9.5
Magna	10	10	10	10	10
Premium	11	11	11	11	11
			Promedio de los promedios =		10.17

3. No hay variación entre poblaciones (tratamientos) pero sí dentro de cada una de ellas.

Los resultados para este caso se muestran en el cuadro 2.4.

Cuadro 2.4

Kilómetros por litro					
Tratamiento	Motor 1	Motor 2	Motor 3	Motor 4	Promedio
Nova	8	10	12	10	10
Magna	10	9	10	11	10
Premium	12	10	8	10	10
			Promedio de los promedios =		10

4. Hay variación entre poblaciones (tratamientos) y también dentro de cada una de las mismas (*véase* el cuadro 2.5).

Cuadro 2.5

Kilómetros por litro					
Tratamiento	Motor 1	Motor 2	Motor 3	Motor 4	Promedio
Nova	9	10	13	10	10.5
Magna	10	9	10	11	10.0
Premium	12	10	11	13	11.5
			Promedio de los promedios =		10.7

Análisis de varianza

El análisis de varianza está fundamentado en una comparación de dos estimaciones diferentes de la varianza $(\sigma)^2$ de una población total.

- La primera consiste en determinar un cálculo de la varianza entre las medias muestrales (medias de los tratamientos).
- La segunda, realizar un cálculo de la varianza dentro de las muestras; es decir, calcular la varianza dentro de cada tratamiento.

Si al comparar ambas estimaciones su valor es aproximadamente igual, se acepta la hipótesis nula (H_0), en caso contrario las diferencias que muestran los diversos tratamientos de un problema son significativas. Para ilustrar ambos procedimientos de cálculo y la prueba de hipótesis correspondiente usaremos el problema siguiente:

Ejemplo 2.1

Una compañía capacita a sus empleados del área de calidad mediante tres técnicas distintas: presencial, a distancia (por internet) y una técnica multimedia. Desea conocer si estos programas representan diversos niveles de productividad para los empleados. Se toma una muestra aleatoria de 14 empleados[3] que han sido capacitados mediante estos tres métodos y se les aplica una evaluación para medir su productividad (*véase* el cuadro 2.6).

Cuadro 2.6 Calificación de empleados.

Técnica	Empleado 1	Empleado 2	Empleado 3	Empleado 4
Presencial	8.5	7.2	8.3	…
A distancia	8.0	8.4	7.8	8.2
Multimedia	8.2	8.0	9.0	8.8

Cálculo de la varianza entre las medias muestrales

Debemos obtener una estimación de la varianza de la población a partir de la varianza entre los *"I"* tratamientos (muestras) que tengamos en el problema. En estadística, esta estimación recibe el nombre de *varianza entre columnas*, ya que la matriz de tratamientos se maneja a través de su transpuesta (*véase* el cuadro 2.7). Para nuestro problema, debemos obtener la varianza entre los tres tratamientos (técnicas de capacitación).

Cuadro 2.7 Técnica de capacitación.

	Presencial	A distancia	Multimedia
Empleado 1	8.5	8.0	8.2
Empleado 2	7.2	8.4	8.0
Empleado 3	8.3	8.1	8.5
Empleado 4	8.0	7.8	9.0
Empleado 5	…	8.2	8.8
Media	8.0	8.1	8.5
Promedio de todas las observaciones			8.2

Para calcular la varianza de cada tratamiento (muestra) empleamos la ecuación 2.1:[4]

$$S^2 = \frac{\sum_{i=1}^{n}(xi - \bar{x})^2}{n-1} \qquad (2.1)$$

Pero nosotros calcularemos la varianza entre tratamientos (entre columnas), por lo que en la ecuación 2.1 debemos sustituir cada dato x_1 por la media de cada tratamiento o muestra $\overline{X}_1$, la media de la muestra $\overline{X}$ por la media de todas las observaciones $\overline{\overline{X}}$ y n el número de datos por k el número de tratamientos o muestras. Con lo cual se obtendrá la ecuación 2.2 de la varianza entre medias muestrales (tratamientos) $S^2_{\bar{X}}$ siguiente:

$$S^2_{\bar{X}} = \frac{\sum_{i=1}^{k}(\bar{X}i - \bar{\bar{X}})^2}{k-1} \qquad (2.2)$$

Con la varianza entre medias muestrales (ecuación 2.2) debemos calcular ahora la varianza entre medias poblacionales a través de la ecuación 2.3 del error estándar de la media:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad (2.3)$$

Elevando al cuadrado (la ecuación 2.3) y despejando σ^2 obtenemos la ecuación de la varianza de la población (2.4).

[3] En la práctica, 14 empleados no constituyen una muestra estadística, pero nos hemos limitado a este número para poder mostrar las técnicas básicas del análisis de varianza, evitando así una gran cantidad de cálculos repetitivos. Por otro lado, es importante recordar que, si los tamaños de las muestras de los tratamientos son suficientemente grandes, no será necesario realizar la suposición de normalidad.

[4] Véase el tema de "Varianza" en Rodríguez J., Pierdant A. y Rodríguez C. (2008), *Estadística para administración*, GE Patria, México, pp. 138-141.

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \qquad\qquad \sigma^2 = n\sigma^2_{\bar{X}} \qquad (2.4)$$

Pero $\sigma^2_{\bar{X}}$ es la varianza entre medias muestrales $S^2_{\bar{X}}$, por lo que al sustituirla en la ecuación 2.4 obtenemos la estimación de la varianza entre columnas (ecuación 2.5) que representa la primera estimación de la varianza de la población con base en la varianza entre medias de tratamientos (muestrales).

$$\sigma^2_{EC} = n\frac{\sum_{i=1}^{K}(\bar{X}_i - \bar{\bar{X}})^2}{k-1} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{\bar{X}})^2}{k-1} \qquad (2.5)$$

donde,

σ^2_{EC} = Primera estimación de la varianza de la población con base en la varianza entre las medias de las muestras o tratamientos (varianza entre columnas).

n_i = Tamaño de la *i-ésima* muestra, $i = 1, 2, 3,\ldots, k$ muestra o tratamiento.

$\bar{X}$ = Media muestral de la *i-ésima* muestra $i = 1, 2, 3,\ldots, k$ muestra o tratamiento.

$\bar{\bar{X}}$ = Gran media es el promedio de todas las observaciones.

k = Número total de muestras o tratamientos.

Para nuestro problema, el cálculo de la estimación de la varianza entre columnas se muestra en el cuadro 2.8.

Cuadro 2.8 Cálculo de la estimación de la varianza entre columnas.

Técnica	n_i	$\bar{X}_i$	$\bar{\bar{X}}$	$\bar{X}_i - \bar{\bar{X}}$	$(\bar{X}_i - \bar{\bar{X}})^2$	$n_i(\bar{X}_i - \bar{\bar{X}})^2$
Presencial	4	8.0	8.2	–0.2	0.040	0.160
A distancia	5	8.1	8.2	–0.1	0.010	0.050
Multimedia	5	8.5	8.2	0.3	0.090	0.450
					SUMA	0.660
Varianza entre columnas = *SUMA*/(*K*–1) = *0.330*						

En la fórmula 2.5, el numerador que representa la suma del cuadrado del valor de todos los tratamientos $\left(SUMA = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2 \right)$ recibe el nombre de suma de cuadrados entre muestras (*SCM*).

Para nuestro problema, $SCM = 0.660$

Cálculo de la varianza dentro de las muestras

El análisis de varianza requiere un segundo cálculo para estimar la varianza de la población. A este cálculo se le denomina *varianza dentro de las muestras* y no es más que un promedio ponderado de la medición de la variación que se presenta dentro de cada muestra o tratamiento. En estadística a este procedimiento se le conoce como varianza dentro de columnas (σ^2_{DC}).

Como indicamos previamente, la varianza de una muestra o tratamiento se puede calcular con la ecuación 2.1.

$$S^2 = \frac{\sum_{i=1}^{n}(xi - \bar{X})^2}{n-1} \qquad (2.1)$$

Dado que un supuesto del análisis de varianza indica que la varianza de las poblaciones de donde se han obtenido las muestras es la misma para todas, esto nos permitiría tomar cualquiera de las varianzas muestrales ($S_1^2,\ S_2^2,\ \ldots\ S_k^2$) como una segunda estimación de la varianza de la población; sin embargo, se puede obtener mejor estimación de la varianza poblacional mediante el promedio ponderado de las varianzas de las *k* muestras. La ecuación 2.6 muestra la forma general para esta segunda estimación de la varianza poblacional σ^2.

$$\sigma_{DC}^2 = \frac{\sum_{i=1}^{k} n_i S_i^2}{\sum_{i=1}^{k} n_i} \qquad (2.6)$$

donde,

σ_{DC}^2 = Segunda estimación de la varianza de la población, con base en la varianza dentro de las muestras (varianza dentro de columnas).

n_i = Tamaño de la *i*-ésima muestra o tratamiento, $i = 1, 2, 3, \ldots, k$ muestra.

k = Número total de muestras o tratamientos.

Para nuestro problema, el cálculo de la varianza dentro de las muestras se efectúa de la manera siguiente:

Calculamos las varianzas de cada muestra o tratamiento (*véase* el cuadro 2.9) con base en la ecuación 2.1.

Cuadro 2.9 Varianzas de cada muestra o tratamiento.

Técnica de capacitación				Varianza 1 $(X_i - \overline{X})^2$	Varianza 2 $(X_i - \overline{X})^2$	Varianza 3 $(X_i - \overline{X})^2$
	Presencial	A distancia	Multimedia			
Empleado 1	8.5	8.0	8.2	0.250	0.010	0.090
Empleado 2	7.2	8.4	8.0	0.640	0.090	0.250
Empleado 3	8.3	8.1	8.5	0.090	0.000	0.000
Empleado 4	8.0	7.8	9.0	0.000	0.090	0.250
Empleado 5	…	8.2	8.8		0.010	0.090
$\overline{X}$	8.0	8.1	8.5			
			Suma =	0.980	0.200	0.680
			Suma/(n –1)	0.327	0.050	0.170

Los cuadrados en cada varianza de las muestras se suman ($Suma = \sum_{i=1} (xi - \overline{X})^2$), como se muestra en el renglón Suma en el cuadro 2.9.

Posteriormente estas sumas se suman nuevamente para obtener la suma de cuadrados dentro de muestras ($SCDM = \sum_{i=1}^{k} Suma_i$).

Para nuestro problema, SCDM = 0.98 + 0.20 + 0.68 = 1.86:

- Calculamos el promedio ponderado de las varianzas de las *k* muestras (*k* = 3 muestras para el problema).

$$\sigma_{DC}^2 = \frac{\sum_{i=1}^{k} n_i s_i^2}{\sum_{i=1}^{k} n_i} \qquad (2.6)$$

$$\sigma_{DC}^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + n_3 s_3^2}{n_1 + n_2 + n_3}$$

$$\sigma_{DC}^2 = \frac{4(0.327)+5(0.050)+5(0.170)}{4+5+5} = \frac{1.308+0.25+0.85}{14}$$
$$\sigma_{DC}^2 = 0.172$$

Con lo cual se obtiene la varianza dentro de las muestras.

Prueba de hipótesis mediante el estadístico *F*

Una vez que contamos con las dos estimaciones de la varianza de la población, el siguiente paso es compararlas mediante el cálculo del cociente siguiente:

$$F = \frac{\textit{Varianza entre las medias muestrales } (\sigma_{EC}^2)}{\textit{Varianza dentro de las muestras } (\sigma_{DC}^2)} \qquad (2.7)$$

Este cociente recibe el nombre de *estadístico F*, en el cual podemos observar que, si ambas varianzas son iguales o parecidas (numerador y denominador), entonces la hipótesis nula es verdadera ($\mu_1 = \mu_2 = \cdots \mu_I$); es decir, la media de las "*k*" poblaciones es igual o casi igual, ya que no hay variabilidad entre muestras ni dentro de las mismas.

En este caso, el estadístico *F* tiende al valor 1 o es muy cercano a éste, por lo que debemos aceptar la hipótesis nula. En caso contrario, conforme el *cociente F* crece, la variabilidad de las medias poblacionales es significativa, por lo cual deberemos aceptar la hipótesis alternativa, las medias poblacionales son significativamente distintas:

$$\mu_1 \neq \mu_2 \neq \cdots \mu_I \text{ (se rechaza la hipótesis nula)}$$

Para nuestro problema, el estadístico *F* toma el valor siguiente:

$$F = \frac{(\sigma_{EC}^2)}{(\sigma_{DC}^2)} = \frac{0.330}{0.172} = 1.918$$

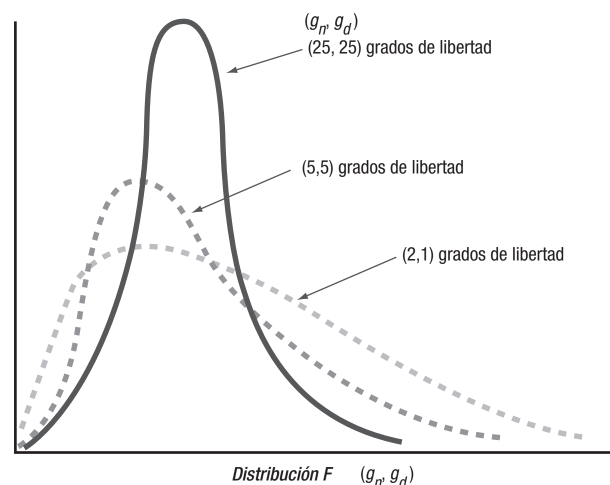
El estadístico *F*, resultado de este cociente, presenta una distribución de probabilidad específica que no es única, ya que cada problema analizado con este método presentaría una distribución de probabilidad distinta; es decir, en realidad es una familia completa de distribuciones (*véase* la gráfica 2.1).

En la gráfica 2.1 se observa que cada distribución está definida por un par de grados de libertad. El primer valor se refiere a los grados de libertad del numerador (g_n) del cociente F; el segundo valor corresponde, a los grados de libertad del denominador (g_d) y que la *distribución F* depende del número de grados de libertad tanto del numerador como del denominador, pero en general está sesgada a la derecha, tiene una sola moda y tiende a ser más simétrica a medida que este par de grados de libertad aumentan.

Ahora bien, para calcular los grados de libertad del numerador (g_n) usaremos la relación siguiente:

$$g_n = (\text{número de muestras} - 1) = (k-1) \qquad (2.7a)$$

Y para calcular los grados de libertad del denominador usamos la relación (2.8):



Gráfica 2.1 Familia completa de distribuciones.

g_d = (total de datos de todas las muestras – el número de muestras)

$$g_d = ((n_1 + n_2 + ... + n_k) - k) \qquad (2.8)$$

Para llevar a cabo la prueba de *hipótesis F* debemos comparar el valor encontrado al comparar las varianzas poblacionales y su valor calculado en una tabla *F* (también podemos usar la *función de distribución inversa de F* de la hoja electrónica de EXCEL =distr.f.inv(alfa, g_n, g_d)).

Si usamos el *cuadro de la distribución de probabilidad F* que se ubica en el anexo de cuadros, debemos especificar primero el nivel de significación alfa (α) que usaremos en la prueba, posteriormente el número de grados de libertad del numerador (g_n), que en el cuadro se ubica en las columnas, y finalmente, los grados de libertad del denominador (g_d), ubicados en los renglones de la tabla.

Para nuestro problema, suponga que el director de capacitación desea probar a un nivel de significancia de 0.05 (α = 0.05) la hipótesis de que no existe diferencia entre las tres técnicas de capacitación. Entonces, una forma de hacerlo es buscar en la *tabla de la distribución F* con α = 0.05 g_n = (3 – 1) = 2, y g_d = (14 – 3) = 11 (*véase* el cuadro 2.10).

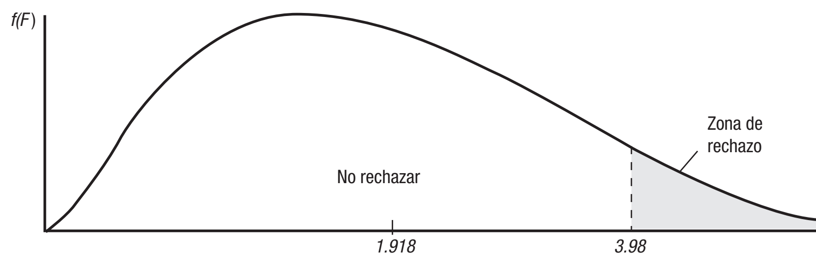
Cuadro 2.10 Distribución *F* con α = 0.05 g_n = (3 – 1) = 2, y g_d = (14 – 3) = 11. *F (0.05,2,11)*

Distribución *F*

$F_{0.95}$; α = 0.05

Grados de libertad del denominador	Grados de libertad del numerador 1	2	3	4	5	6	7	8	9
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65

El valor encontrado en la tabla es de 3.98 y establece el límite superior de la región de aceptación, como se muestra en la gráfica 2.2.



Gráfica 2.2 Límite superior de la región de aceptación.

Como el valor de la muestra calculado, *F* = 1.918, se encuentra dentro de la región de aceptación, aceptamos la hipótesis nula y concluimos que, según la información estadística de las muestras que poseemos, no existen diferencias significativas en su nivel de productividad al recibir una capacitación en calidad por cualquiera de estas tres técnicas.

En Excel, el valor de F se obtiene con la función:

$$=\text{DISTR.F.INV}(\alpha, g_n, g_d)$$

Sustituyendo valores:

$$=\text{DISTR.F.INV}(0.05, 2, 11)$$

Obtenemos:

$$3.98229796$$

Cuadro resumen del análisis de varianza para un factor

En el ámbito estadístico se acostumbra elaborar un cuadro resumen de los cálculos realizados en un análisis de varianza (*véase* el cuadro 2.11).

Cuadro 2.11 Formato del cuadro para el análisis de varianza de un factor de muestra.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	F	p
Entre muestras (tratamientos)	SCM	$k-1$	σ^2_{EC}	$\frac{(\sigma^2_{EC})}{(\sigma^2_{DC})}$	Significancia (sig.)
Dentro de muestras de error	SCDM	$n-k$	σ^2_{DC}		
Variación total	SCM + SCDM	$n-1$			

Nota: $n = n_1 + n_2 + \ldots + n_k$, suma del número de datos de todas las muestras.
k = número de muestras o tratamientos.

Para nuestro problema, el cuadro resumen del análisis de varianza para el factor tipo de capacitación (*véase* el cuadro 2.12) es:

Cuadro 2.12 Resumen del análisis de varianza para el factor tipo de capacitación.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	F	p
Entre muestras (tratamientos)	0.66	2	0.33	0.192	0.189
Dentro de muestras de error	1.86	11	0.172		
Variación total	2.52	13			

Mediante el paquete estadístico SPSS para el problema, el cuadro de análisis de varianza de un factor se muestra en el cuadro 2.13.

Cuadro 2.13 Análisis de varianza de un factor.

ANOVA					
Calificación en evaluación					
	Sume of squares	*df*	*Mean square*	*F*	*Sig.*
Between groups	0.657	2	0.329	1.943	0.189
Within groups	1.860	11	0.169		
Total	2.517	13			

Empleo del valor *p* en las pruebas de hipótesis[5]

Con el uso de los paquetes estadísticos para computadoras, apareció el concepto del *valor p*, un nuevo enfoque que permite probar hipótesis.

El *valor p* es la probabilidad de obtener un estadístico de prueba igual o más extremo que el resultado obtenido a partir de los datos muestrales, dado que la hipótesis nula H_0 es cierta.

En otras palabras, el *valor p* es el nivel más bajo de significancia (α) al cual se puede rechazar la hipótesis nula. Comprende el área en la cola que está más allá del valor del estadístico para la muestra. A medida que el estadístico de la prueba se adentra en la región de rechazo, indica mayor evidencia para rechazar la hipótesis nula, observando que el *valor p* se hace más pequeño.

Suponiendo que la hipótesis nula H_0 es cierta, un *valor p* muy pequeño es una fuerte evidencia para rechazar la hipótesis nula, ya que indica que el dato observado es muy poco probable que se presente.

Al *valor p* también se le conoce como el *nivel observado de significancia*.

La regla, por tanto, para rechazar una hipótesis nula mediante el uso del *valor p* es:

$$\text{Rechace } H_0 \text{ si el valor } p < \alpha$$

Independientemente del tipo de prueba de hipótesis a que se haga referencia, esta regla es válida para pruebas de hipótesis de dos extremos o bien de un extremo.

Para nuestro problema, el *valor p* es mayor que α (0.189 > 0.05), por lo que debe aceptarse la hipótesis nula.

Ejemplo 2.2

La Secretaría de Ecología del gobierno de Nuevo León está analizando establecer nuevas reglas ecológicas debido a la concentración de contaminantes en la ciudad de Monterrey. La secretaría propone un reglamento que impone reglas específicas de control de contaminantes en cada estación del año, bajo la hipótesis de que las condiciones ambientales cambian esta concentración. El ayuntamiento considera que esto no es necesario, pues su hipótesis versa en el sentido de que la concentración de contaminantes es similar independientemente de la estación del año. Por tanto, se solicita el estudio a la universidad estatal con la finalidad de determinar la mejor política pública para el control de contaminantes en la ciudad.

[5] Tomado de Rodríguez J., Pierdant A. y Rodríguez C. (2008), *Estadística para administración*, Grupo Editorial Patria, México, pp. 401-402.

El laboratorio de ecología de la Universidad Autónoma de Nuevo León (UANL) cuenta con una muestra de seis mediciones realizadas en el transcurso de cada una de las cuatro estaciones del año pasado (*véase* el cuadro 2.14).

Cuadro 2.14 Concentración de contaminantes en Monterrey, Nuevo León.

Medición	Primavera	Verano	Otoño	Invierno
1	5.62	7.70	2.52	6.77
2	6.12	8.31	5.44	6.65
3	6.62	8.80	4.94	6.01
4	6.21	8.24	2.99	6.26
5	7.08	7.87	4.39	7.09
6	5.36	7.44	4.44	6.05

La universidad decide aplicar un análisis de varianza para el factor *estación del año*, y probar la hipótesis del ayuntamiento en el sentido de que la concentración de contaminantes no cambia con la estación del año.

- $H_0: \mu_1 = \mu_2 = ... = \mu_I$; es decir, la media de concentración de contaminantes es igual en cualquier estación del año (hipótesis del ayuntamiento de Monterrey).
- $H_1: \mu_1 \neq \mu_2 \neq ... \neq \mu_I$; es decir, las medias de concentración de contaminantes son distintas y dependen de las características de cada estación del año (hipótesis de la Secretaría de Ecología estatal).

Solución

Una vez que hemos establecido ambas hipótesis, el primer paso consiste en determinar la *varianza entre las medias muestrales* o *varianza entre columnas* (*véase* el cuadro 2.15).

Cuadro 2.15 Varianza entre columnas.

Estación	n_i	$\overline{X}$	$\overline{\overline{X}}$	$\overline{X} - \overline{\overline{X}}$	$(\overline{X} - \overline{\overline{X}})_2$	$n_i(\overline{x} - \overline{\overline{x}})^2$
Primavera	6	6.2	6.2	0.0	0.001	0.008
Verano	6	8.1	6.2	1.9	3.460	20.758
Otoño	6	4.1	6.2	–2.1	4.326	25.958
Invierno	6	6.5	6.2	0.3	0.074	0.443
					Suma =	47.167
Varianza entre columnas = Suma/ (k – 1) = 15.722						

Por tanto, del cuadro 2.15 sabemos que SCM = 47.167.

En este método se requiere un segundo cálculo para estimar la varianza de la población denominado la *varianza dentro de las muestras* (*véase* el cuadro 2.16).

Cuadro 2.16 Varianza dentro de las muestras.

	Estación del año				Varianza 1	Varianza 2	Varianza 3	Varianza 4
medición	Primavera	Verano	Otoño	Invierno	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2$
1	5.62	7.7	2.52	6.77	0.301	0.130	2.560	0.089
2	6.12	8.31	5.44	6.65	0.002	0.063	1.742	0.032
3	6.62	8.8	4.94	6.01	0.204	0.548	0.672	0.213
4	6.21	8.24	2.99	6.26	0.002	0.032	1.277	0.045
5	7.08	7.87	4.39	7.09	0.831	0.036	0.073	0.382
6	5.36	7.44	4.44	6.05	0.653	0.384	0.102	0.178
$\overline{X}$	6.2	8.1	4.1	6.5				
				Suma =	1.993	1.193	6.427	0.939
				Suma/(n-1)	0.399	0.239	1.285	0.188

- Calculamos la suma de cuadrados dentro de la media $\left(SCDM = \sum_{i=1}^{k}(Suma)_i\right)$

$$\text{SCDM} = 1.993 + 1.193 + 6.427 + 0.939 = 10.552$$

- Calculamos el promedio ponderado de las varianzas de las k muestras ($k = 4$) para obtener la varianza dentro de las muestras.

$$\sigma^2_{DC} = \frac{\sum_{i=1}^{k} n_i s_i^2}{\sum_{i=1}^{k} n_i} \qquad (2.6)$$

$$\sigma^2_{DC} = \frac{6(0.399) + 6(0.239) + 6(1.285) + 6(0.188)}{6+6+6+6} = \frac{12.662}{24} = 0.528$$

- Con las dos estimaciones de la varianza de la población, calculamos el *estadístico F.*

$$F = \frac{\left(\sigma^2_{EC}\right)}{\left(\sigma^2_{DC}\right)}$$

$$F = \frac{15.722}{0.528} = 29.776$$

Determinamos los grados de libertad tanto del numerador (g_n) como del denominador (g_d) y obtenemos el *valor del estadístico F* en los cuadros o mediante Excel.

$$g_n = (\text{número de muestra} - 1) = (4 - 1) = 3$$

$$g_d = (\text{total de datos de todas las muestras} - \text{el número de muestras})$$

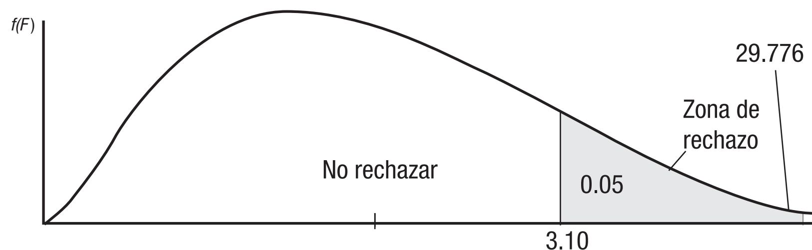
$$g_d = ((n_1 + n_2 + \ldots n_k) - k) = 24 - 4 = 20$$

En el ejemplo, el ayuntamiento de Monterrey desea probar que la concentración de contaminantes es igual en cada estación del año, por lo que propone probar su hipótesis a un nivel de significancia de 0.05 ($\alpha = 0.05$). Entonces, una forma de hacerlo es buscar el valor en la tabla de la distribución F con $\alpha = 0.05$, $g_n = 3$, y $g_d = 20$.

- El valor en los cuadros encontrado es: 3.10
- con Excel, =DISTR.F.INV(0.05, 3, 20)
  Obtenemos:

3.0984

- Este valor establece el límite superior de la región de aceptación, como se muestra en la gráfica 2.3.



Gráfica 2.3 Límite superior de la región de aceptación.

- Como el valor calculado de $F = 29.776$, éste se encuentra fuera de la región de aceptación ($29.776 > 3.10$), rechazamos la hipótesis nula y concluimos que, según la información estadística de las muestras que poseemos, existen diferencias significativas en la concentración de contaminantes durante las estaciones del año en la ciudad.
- Entonces deberá aceptarse la propuesta de la Secretaría de Ecología estatal.
- Con la información de los cálculos realizados en nuestro ejemplo, podemos elaborar un cuadro resumen del análisis de varianza del factor: estación del año (*véase* el cuadro 2.17).

Cuadro 2.17

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	*F*	*p*
Entre muestras (tratamiento)	47.167	3	15.722	29.776	0.000
Dentro de muestras (error)	10.552	20	0.528		
Variación total	57.719	23			

- En el cuadro 2.17 se observa que el valor p es menor que α ($0.00 < 0.05$), lo que nos indica que debemos rechazar la hipótesis nula (H_0) y aceptar la hipótesis alternativa (H_1).
- Por tanto, se comprueba nuevamente que deberá aceptarse la propuesta de la Secretaría de Ecología estatal.

Análisis de varianza para un factor con SPSS

Un análisis de varianza para un factor mediante el uso del paquete Statistical Package for Social Sciences (SPSS), versión 16.0 para Windows puede elaborarse realizando los pasos siguientes.

1. Primero se definen las variables de análisis.
   Retomando el ejemplo anterior, las variables de análisis quedarían definidas en la ventana de vista de variables (*Variable view*) como Estación (1, Primavera, 2, Verano, 3, Otoño y 4, Invierno) y Concentración (*véase* la figura 2.1 del editor de SPSS).
2. Se capturan los datos en la ventana de vista de datos, Data View (*véase* la figura 2.2).

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	ESTACIÓN	Numeric	1	0	Estación del año	{1, Primavera}...	None	8	Center	Scale
2	CONCENTRACIÓN	Numeric	5	2	Concentración de contaminantes	None	None	8	Center	Scale

Figura 2.1 Variables de análisis definidas en la ventana de vista de variables (Variable view).

	ESTACIÓN	CONCENTRACIÓN
1	1	5.62
2	1	6.12
3	1	6.62
4	1	6.21
5	1	7.08
6	1	5.36
7	2	7.70
8	2	8.31
9	2	8.80
10	2	8.24
11	2	7.87
12	2	7.44
13	3	2.52
14	3	5.44
15	3	4.94
16	3	2.99
17	3	4.39
18	3	4.44
19	4	6.77
20	4	6.65
21	4	6.01
22	4	6.26
23	4	7.09
24	4	6.06

Figura 2.2 Ventana de datos, Data View.

3. Se solicita el cálculo del análisis de varianza para un factor, por tanto, es necesario del menú **Analize**, submenú, **Compare means**, seleccionamos la rutina **One-Way ANOVA**… (*véase* la figura 2.3).

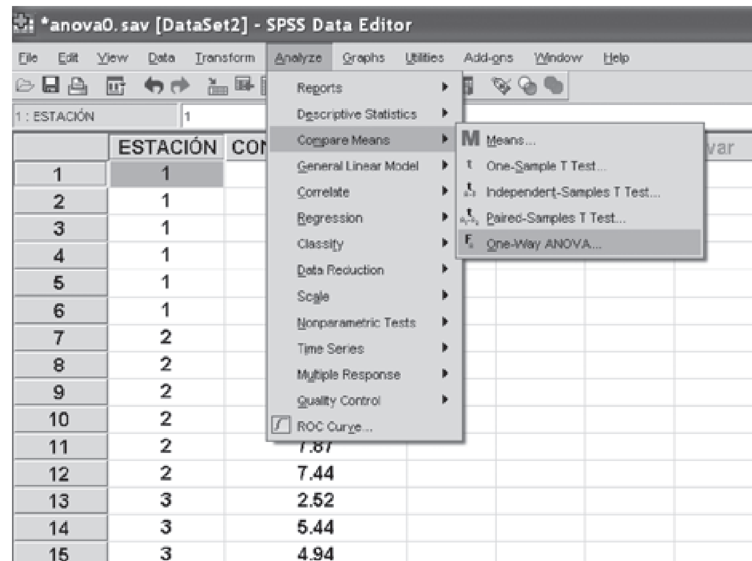


Figura 2.3 Cálculo del análisis de varianza para un factor.

4. Seleccionamos la variable dependiente (concentración de contaminantes) y la variable independiente o factor (estación).

   En la sección de opciones (**Options**…) solicitar las estadísticas descriptivas de la variable dependiente (*véase* la figura 2.4).

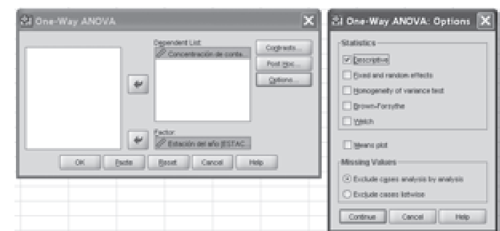


Figura 2.4 En la sección de opciones (Options…) solicitar las estadísticas descriptivas de la variable dependiente.

5. Solicitar el cálculo mediante OK en la ventana de **One-Way Anova** para obtener los resultados que se muestran en la figura 2.5.

Oneway

(Processing...)

[DataSet2] C : \ Documents and Settings \ Alberto Pierdant \ Mis documentos \ A.

Descriptives

Concentración de contaminantes

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Primavera	6	6.1683	.63139	.25776	5.5057	6.8309	5.36	7.08
Verano	6	8.0600	.48839	.19938	7.5475	8.5725	7.44	8.80
Otoño	6	4.1200	1.13375	.46285	2.9302	5.3098	2.52	5.44
Invierno	6	6.4717	.43333	.17691	6.0169	6.9264	6.01	7.09
Total	24	6.2050	1.58411	.32335	5.5361	6.8739	2.52	8.80

ANOVA

Concentración de contaminantes

	Sum of Squares	df	Mean Square	F	Big
BetweenGroups	47.164	3	15.721	29.799	.000
Within Groups	10.552	20	.528		
Total	57.716	23			

Figura 2.5 Cálculo mediante OK en la ventana de One-Way Anova.

6. A través de este simple procedimiento de cálculo podemos obtener un análisis de varianza para un factor con el paquete estadístico SPSS.

Pruebas para la diferencia entre pares de medias

En el análisis de varianza de un factor el resultado de aceptar la hipótesis nula (H_0) nos puede indicar si todas las medias en el análisis son iguales; sin embargo, cuando se rechaza la hipótesis nula, este análisis no nos indica cuál o cuáles medias son diferentes entre sí.

Para determinar la diferencia entre dos medias requerimos otras pruebas estadísticas, las cuales consisten en realizar una comparación por pares, de todos los pares de medias posibles.

Si cualesquier valor absoluto de la diferencia entre dos medias muestrales es mayor que algún criterio de comparación, entonces se dice, que esta diferencia es significativa y se concluye que las medias poblacionales respectivas son diferentes.

En estadística existen diversos criterios de comparación de pares de medias, pero en la práctica los más utilizados son el método de Tukey (léase *TooKey*) y el de diferencia mínima significativa (DMS).

Ambos métodos se pueden utilizar indistintamente si el número de observaciones en cada muestra o tratamiento es el mismo ($n_1 = n_2 = n_3 = \ldots n_k$). En estos casos, se dice que estos problemas de análisis presentan un diseño de análisis de varianza balanceado.

Si el problema de estudio presenta muestras con tamaños diferentes ($n_1 \neq n_2 \neq n_3 \neq \ldots \neq n_k$), se dice que estos problemas de análisis de varianza presentan un diseño no balanceado y podrán analizarse a través de un método DMS modificado.

Prueba de Tukey y de DMS para diseños balanceados

Como indicamos previamente, un diseño balanceado en un análisis de varianza indica que todas las muestras del análisis tienen el mismo número de observaciones. En estos casos el analista puede emplear la prueba de Tukey o bien la de DMS. Para ilustrar el procedimiento de ambas pruebas utilizaremos el problema 2.1.

Ejemplo 2.3

La cadena de restaurantes VIKS cuenta con cuatro unidades en el sur de la ciudad. El gerente de la zona desea realizar una campaña de promoción de juegos y premios entre los comensales para incrementar las ventas. El gerente considera que diferentes juegos y premios atraería a diferentes tipos de consumidores, desde los que tienen ingresos medios hasta los que tienen ingresos altos. Decide utilizar el monto de los consumos como medida representativa del ingreso. Desea determinar si existe diferencia en el nivel promedio de consumo entre las cuatro unidades de su zona. Si encuentra alguna diferencia, entonces, ofrecerá mayor diversidad de premios en su campaña de promoción.

Cuadro 2.18 Resumen de datos y las medias de consumo.

Nota de consumo	San Ángel	Tláhuac	Taxqueña	Culhuacán
1	510	190	360	130
2	490	190	420	150
3	560	210	450	90
4	480	240	480	100
5	380	210	390	190
6	510	310	410	150
7	480	250	510	210
Medias	487.14	228.57	431.43	145.71

Nota: El total de la nota está en pesos.

- El cuadro 2.19 muestra un resumen del análisis de varianza para un factor de este problema.

Cuadro 2.19 Anova, resumen del análisis de varianza para un factor de este problema.

Consumo	*Sum of squares*	*df*	*Mean square*	*F*	*Sig.*
Between groups	553325.000	3	184441.667	78.090	0.000
Within groups	56685.714	24	2361.905		
Total	610010.714	27			

En el cuadro 2.19 el valor p es menor que el valor de α ($0.000 < 0.05$) por lo que debemos rechazar la hipótesis nula (H_0), es decir, los consumos promedio por restaurante son distintos. El siguiente paso consiste en determinar cuáles son diferentes.

Prueba de Tukey para diseños balanceados

Este método se desarrolló en 1953 por J.W. Tukey y requiere el cálculo del criterio de T. Tukey, definido como:

$$T = q_{\alpha,\,k,\,n-k}\sqrt{\frac{\sigma^2_{DC}}{r}} \qquad (2.9)$$

donde,

q = distribución de rangos elaborada a través de una t de Student con k (tratamientos) y ($n - k$) grados de libertad para un valor α.

k = número de muestras o tratamientos.

n = total de datos de todas las muestras.

r = número de datos por muestra o tratamiento (r debe ser igual en todas las muestras).

σ^2_{DC} = la varianza dentro de muestras (error).

- El valor $q_{\alpha,\,k,\,n-k}$ para un determinado valor de α se puede obtener de la tabla de valores críticos de la distribución de rangos de Student que se ubica en el anexo de tablas.
- Si el nivel de significancia deseado es del 5%, entonces el valor q para nuestro problema toma el valor siguiente:

$$q_{0.05,\,4,\,24} = 3.9$$

- Del cuadro resumen del análisis de varianza para un factor 2.19 obtenemos el valor de la varianza dentro de las muestras:

$$\sigma^2_{DC} = 2361.905$$

- Sustituyendo valores en la ecuación del criterio de Tukey obtenemos:

$$T = q_{\alpha,\, k,\, n-k}\sqrt{\frac{\sigma^2_{DC}}{r}} = 3.9\sqrt{\frac{2361.905}{7}} = 71.64$$

- El criterio T se compara con la diferencia absoluta entre cada par de medias muestrales (*véase* el cuadro 2.20).

Cuadro 2.20 Criterio *T* se compara con la diferencia absoluta entre cada par de medias muestrales.

SA -Tla	| 487.14 – 228.57 |	=	258.57	>	71.64
SA -Tax	| 487.14 – 431.43 |	=	55.71	<	71.64
SA - Cul	| 487.14 – 145.71 |	=	341.43	>	71.64
Tla - Tax	| 228.57 – 431.43 |	=	202.86	>	71.64
Tla - Cul	| 228.57 – 145.71 |	=	82.86	>	71.64
Tax - Cul	| 431.43 – 145.71 |	=	285.71	>	71.64

- Se observa que sólo las unidades San Ángel y Taxqueña tienen igual nivel de consumo promedio y las otras diferencias calculadas exceden el criterio T.
- En el cuadro 2.21 se muestra la salida de este método mediante el paquete SPSS.

Cuadro 2.21 Salida del criterio *T* mediante el sistema SPSS, con múltiples comparaciones.

Consumo Tukey HSD

(I) Restaurante	*(J) Restaurante*	*Mean difference (I-J)*	*Std error*	*Sig.*	*95% Intervalo confidencial*	
					Lower bound	*Upper bound*
San Ángel	Tláhuac	258.571*	25.977	0.000	186.91	330.23
	Taxqueña	55.714	25.977	0.168	–15.95	127.38
	Culhuacán	341.429*	25.977	0.000	269.77	413.09
Tláhuac	San Ángel	–258.571*	25.977	0.000	–330.23	–186.91
	Taxqueña	–202.857*	25.977	0.000	–274.52	–131.20
	Culhuacán	82.857*	25.977	0.019	11.20	154.52
Taxqueña	San Ángel	–55.714	25.977	0.168	–127.38	15.95
	Tláhuac	202.857*	25.977	0.000	131.20	274.52
	Culhuacán	285.714*	25.977	0.000	214.05	357.38
Culhuacán	San Ángel	–341.429*	25.977	0.000	–413.09	–269.77
	Tláhuac	–82.85*	25.977	0.019	–154.52	–11.20
	Taxqueña	–285.714*	25.977	0.000	–357.38	–214.05

* The mean difference is significant at the 0.05 level

- En el cuadro 2.21 se comprueba en la segunda columna (*mean difference*) que sólo las unidades San Ángel y Taxqueña no tienen una diferencia significativa en los consumos promedio de sus clientes a un nivel de significancia de 0.05.
- Lo mismo lo comprobamos nuevamente, ya que el valor p es mayor que α $(0.168 > 0.05)$; es decir, se acepta la hipótesis nula (*medias iguales*).

Prueba de la diferencia mínima significativa (DMS) para diseños balanceados

El método de la diferencia mínima significativa (DMS) es similar al método de Tukey, ya que compara el criterio DMS con la diferencia absoluta calculada con las medias muestrales.

- Si el diseño está balanceado, el criterio DMS es:

$$DMS = \sqrt{\frac{2\sigma^2_{DC} F_{\alpha,1,n-k}}{r}} \qquad (2.10)$$

donde,

F = Distribución F con 1 y $n-k$ grados de libertad para un valor α.
k = Número de muestras o tratamientos.
n = Total de datos de todas las muestras.
r = Número de datos por muestra o tratamiento (r debe ser igual en todas las muestras).
σ^2_{DC} = La varianza dentro de muestras (error).

- De las tablas o con EXCEL el valor de $F_{0.05,\,1,\,24}$ es de 4.26.
- Sustituyendo valores en la ecuación 2.10 obtenemos:

$$DMS = \sqrt{\frac{2\sigma^2_{DC} F_{\alpha,\,1,\,n-k}}{r}} = \sqrt{\frac{2(2361.9050)(4.26)}{7}} = 53.62$$

- El criterio se compara con la diferencia absoluta entre cada par de medias muestrales, al igual que el criterio T (*véase* el cuadro 2.22).

Cuadro 2.22 Diferencia absoluta entre cada par de medias muestrales.

SA -Tla	| 487.14 – 228.57 |	=	258.57	>	53.62
SA -Tax	| 487.14 – 431.43 |	=	55.71	>	53.62
SA - Cul	| 487.14 – 145.71 |	=	341.43	>	53.62
Tla - Tax	| 228.57 – 431.43 |	=	202.86	>	53.62
Tla - Cul	| 228.57 – 145.71 |	=	82.86	>	53.62
Tax - Cul	| 431.43 – 145.71 |	=	285.71	>	53.62

- Se observa en el cuadro 2.22 que ninguna unidad tiene igual nivel de consumo promedio que otra, bajo este criterio. Todas las diferencias calculadas exceden el criterio DMS.
- El criterio DMS es más conservador que el valor Tukey.
- Y al igual que el criterio anterior se puede calcular usando el paquete estadístico SPSS (*véase* el cuadro 2.23).

Cuadro 2.23 Consumo Tukey HSD.

Multiple Comparisons

Consumo
LSD

(I) Restaurante	(J) Restaurante	Mean difference (I-J)	Std error	Sig.	95% Intervalo confidencial	
					Lower bound	Upper bound
San Ángel	Tláhuac	258.571*	25.977	.000	204.96	312.19
	Taxqueña	55.714*	25.977	.042	2.10	109.33
	Culhuacán	341.429*	25.977	.000	287.81	395.04
Tláhuac	San Ángel	–258.571*	25.977	.000	–312.19	–204.96
	Taxqueña	–202.857*	25.977	.000	–256.47	–149.24
	Culhuacán	82.857*	25.977	.004	29.24	136.47
Taxqueña	San Ángel	–55.714*	25.977	.042	–109.33	–2.10
	Tláhuac	202.857*	25.977	.000	149.24	256.47
	Culhuacán	285.714*	25.977	.000	232.10	339.33
Culhuacán	San Ángel	–341.429*	25.977	.000	–395.04	–287.81
	Tláhuac	–82.857*	25.977	.004	–136.47	–29.24
	Taxqueña	–285.714*	25.977	.000	–339.33	–232.10

* The mean difference is significant at the 0.05 level

Prueba DMS modificada para diseños no balanceados

Si el problema de estudio presenta muestras con tamaños diferentes $[(n)]_1 \neq n_2 \neq n_3 \neq ... \neq n_k)$, se dice que estos problemas de análisis de varianza presentan un diseño no balanceado y únicamente podrán analizarse mediante un método DMS modificado.

- En el método DMS modificado para comparar las muestras *i*-ésima y *j*-ésima, el criterio de comparación DMS se modifica a:

$$DMS_{i,j} = \sqrt{\left[\frac{1}{r_i} + \frac{1}{r_j}\right] \sigma^2_{DC} F_{\alpha,\, k-1,\, n-k}} \qquad (2.11)$$

donde,

F = Distribución F con k–1 y n–k grados de libertad para un valor α.

k = Número de muestras o tratamientos.

n = Total de datos de todas las muestras.

r_i = Número de datos para la muestra o tratamiento i-ésima.

r_j = Número de datos para la muestra o tratamiento *j*-ésima.

σ^2_{DC} = La varianza dentro de muestras (error).

- El criterio DMS modificado será diferente para cada par de comparaciones de medias, debido a que el número de observaciones no es el mismo en cada muestra.

Ejemplo 2.4

Suponga un análisis de varianza para tres tratamientos cuyas muestras tienen diferente tamaño (*véase* el cuadro 2.24). Determine si las muestras tienen la misma media. Si no tienen la misma media cuáles pares son iguales.

Cuadro 2.24

	Tratamientos		
Datos	I	II	III
1	30	38	19
2	25	32	35
3	31	35	20
4	35	36	22
5		38	25
6		32	
Medias	30.25	35.17	24.20

Solución

- Suponiendo que $\alpha = 0.05$, entonces de tablas o con EXCEL, $F_{0.05,(3-1),(15-3)} = 3.89$. La salida **One-Way ANOVA** del paquete SPSS se muestra en el cuadro 2.25.

Cuadro 2.25 Salida **One-Way ANOVA** del paquete SPSS.

Datos	*Sum of squares*	*df*	*Mean square*	*F*	*Sig.*
Between groups	328.017	2	164.008	7.737	0.007
Within groups	254.383	12	21.199		
Total	582.400	14			

- Puesto que la *F* calculada (7.737) es mayor que la *F* de tablas (3.89), debemos rechazar la hipótesis nula, es decir, las medias de los tratamientos son distintas. Este resultado se confirma a través del valor *p*, ya que *p* es menor que α ($0.007 < 0.05$).
- Como las medias de los tratamientos son diferentes, entonces nos resta analizar que pares de ellas son iguales (*I-II*, *I-III* y *II-III*). Pero observamos que el número de datos "*r*" en cada una de ellas es distinto por lo que deberemos aplicar una prueba DMS modificada para diseños no balanceados.
- Si $\alpha = 0.05$ y $F_{0.05,(3-1),(15-3)} = 3.89$, la comparación I-II para DMS modificada es:

$$DMS_{I,\,II} = \sqrt{\left[\frac{1}{4}+\frac{1}{6}\right](21.199)(3.89)} = 5.86$$

- La comparación *I-III* para DMS modificada es:

$$DMS_{I,\,III} = \sqrt{\left[\frac{1}{4}+\frac{1}{5}\right](21.199)(3.89)} = 6.09$$

- La comparación *II-III* para DMS modificada es:

$$DMS_{II,\,III} = \sqrt{\left[\frac{1}{6}+\frac{1}{5}\right](21.199)(3.89)} = 5.49$$

- La diferencia entre medias comparada con su DMS modificado se muestra en el cuadro 2.26.

Cuadro 2.26 Diferencia entre medias comparada con su DMS modificado.

I - II	| 30.25 – 35.17 |	=	4.92	<	5.86
I - III	| 30.25 – 24.20 |	=	6.05	<	6.09
II - III	| 35.17 – 24.20 |	=	10.97	>	5.49

- Sólo el tratamiento *II* y *III* difieren significativamente, el resultado también se puede obtener mediante el paquete SPSS (*véase* el cuadro 2.27).

Cuadro 2.27 Los tratamientos *II* y *III* difieren significativamente.

Datos LSD					*95% Confidence interval*	
(I)Tratamiento	*(II)Tratamiento*	*Mean difference (I–J)*	*Std. error*	*Sig.*	*Lower bound*	*Uper bound*
I	II	–4.91667	2.97200	0.124	–11.3921	1.5588
	III	6.05000	3.08859	0.074	–0.6795	12.7795
II	I	4.91667	2.97200	0.124	–1.5588	11.3921
	III	10.96667*	2.78798	0.002	4.8922	17.0411
III	I	–6.05000	3.08859	0.074	–12.7795	0.6795
	II	–10.96667*	2.78798	0.002	–17.0411	–4.8922

Análisis de varianza con dos factores

En el análisis de varianza para un factor se considera que sólo un factor influye en las unidades experimentales; sin embargo, con frecuencia se observa que un segundo factor exterior puede influenciar el comportamiento de estas unidades experimentales.

Un análisis estadístico que considera simultáneamente ambos factores recibe el nombre de análisis de varianza con dos factores (**Two-Way ANOVA**).

Por ejemplo, el Gobierno del Distrito Federal cuenta con tres tipos de máquinas para colocar el asfalto en las calles de la ciudad.

Se desea comparar la productividad promedio (km/día) de las tres máquinas de asfalto (tratamientos); sin embargo, el ingeniero de campo se da cuenta que al probarlas la destreza del operador y su experiencia pueden afectar el número de kilómetros por día asfaltados, lo que produce una confusión sobre cuál máquina es realmente mejor.

Para obtener un panorama no contaminado y más claro sobre la productividad de cada máquina se debe eliminar o corregir de alguna manera la influencia del operador sobre la productividad final de cada equipo. Esta consideración simultánea de dos factores (máquina y operador) requiere un análisis de varianza con dos factores.

Para solucionar este problema y realmente obtener una medida de la capacidad de poner asfalto de estas máquinas, debemos *bloquear* el factor externo (el operador), colocando las observaciones en grupos homogéneos con base, por ejemplo, en los años de experiencia que tiene cada operador (la máquina de asfalto 1 con los operadores que tienen 1 año de experiencia, otro con los que tienen 2 años, y así sucesivamente. Por tanto, las observaciones se clasifican por bloques (años de experiencia) y por tratamientos (máquinas). El propósito del bloqueo es reducir la variación dentro de un tratamiento. A este diseño experimental se le conoce como *diseño aleatorizado en bloques.*

Si los bloques se elaboran adecuadamente con base en un factor (la experiencia) que verdaderamente afecte la productividad se obtiene una medición más precisa del efecto del tratamiento; sin embargo, si el factor seleccionado para el bloqueo no es el adecuado, los resultados pueden ser engañosos; por tanto, es importante seleccionar adecuadamente el factor de bloqueo para garantizar que sí tenga cierto impacto.

Para mostrar el procedimiento de cálculo de un análisis de varianza con dos factores usaremos el problema 2.3.

Ejemplo 2.5

El Banco de Comercio trata de seleccionar tres nuevos sistemas de cómputo para mejorar la calidad de la atención de sus clientes. La selección final del sistema dependerá de su productividad (clientes atendidos por hora). Por tanto, se seleccionan aleatoriamente cinco cajeros para cada sistema y los tres proveedores de los sistemas le indican al banco que, para un manejo adecuado de sus sistemas considere la experiencia de sus cajeros en la prueba del sistema, ya que este factor puede afectar el resultado en la productividad de su respectivo sistema.

La gerencia de sistemas desea evaluar el impacto de la experiencia de los cajeros en la selección del sistema de atención de clientes. En una prueba de los sistemas, el número de clientes por hora en cada sistema se muestra en el cuadro 2.28.

Cuadro 2.28 Evaluación del impacto de la experiencia de los cajeros en la selección del sistema de atención de clientes.

	Sistema de atención a clientes (tratamiento)		
Años de experiencia del cajero	Oracle (clientes/hora)	Total (clientes/hora)	Adabas (clientes/hora)
1	25	21	27
2	35	33	31
3	39	39	42
4	37	41	38
5	45	46	45

Dentro de una muestra (tratamiento) dada se observará una variación en la productividad (clientes/hora) debido a la experiencia del cajero, la capacitación recibida y otros factores de error aleatorios. Pero el banco, no está interesado en la productividad de los cajeros, sino en la productividad del sistema de atención a clientes, por lo que se debe ajustar la productividad del cajero para eliminar su efecto de variabilidad y obtener así, una medida precisa de la calidad del sistema.

Análisis de varianza con dos factores (diseño aleatorizado en bloques)

En el análisis de varianza con dos factores, la suma de cuadrados total se divide en tres partes:

- La suma de cuadrados entre muestras (tratamientos) (SCM).
- La suma de cuadrados dentro de las muestras (SCDM).
- Suma de cuadrados de bloques (SCB).

Por tanto la suma de cuadrados total (SCT) será:

$$\text{SCT} = \text{SCM} + \text{SCDM} + \text{SCB} \qquad (2.12)$$

La suma de cuadrados total (SCT) y la suma de cuadrados entre muestras (SCM) se calculan de la misma forma que en el análisis de varianza para un factor; sin embargo, la suma de cuadrados dentro de las muestras (SCDM) se subdivide en una medida para SCDM y otra para la suma de cuadrados de bloques (SCB).

La suma de cuadrados de bloques se define como:

$$SCB = \sum_{i=1}^{k} k\left(\bar{X}_i - \bar{\bar{X}}\right)^2 \qquad (2.13)$$

donde,

k = número de muestras o tratamientos:

$\bar{X}_1$ = medio del i-ésimo bloque $i = 1, 2, 3, ..., b$.

$\bar{\bar{X}}$ = la gran media.

Para nuestro problema de **ANOVA** con dos factores, el cálculo de la suma de cuadrados de bloques se muestra en el cuadro 2.29.

Cuadro 2.29 ANOVA con dos factores.

	Sistema de atención a clientes (tratamiento)				
Años de experiencia del cajero	Oracle (clientes /hora)	Total (clientes /hora)	Adabas (clientes/hora)	x_i	$k(\bar{r}-\bar{\bar{r}})^2$
1	25	21	27	24.33	427.21
2	35	33	31	33.00	32.01
3	39	39	42	40.00	41.81
4	37	41	38	38.67	17.28
5	45	46	45	45.33	246.61
	$\bar{\bar{x}}$	36.267		SCB =	764.93

- La suma de cuadrados de bloques mide el grado de variación de las medias del bloque (filas) alrededor de la gran media.
- La suma de cuadrados entre muestras (SCM) se calcula con el numerador de la ecuación (2.5):

$$SCM = \sum_{j=1}^{b} n_j\left(\bar{X}_j - \bar{\bar{X}}\right)^2 \qquad (2.5)$$

- Los resultados del cálculo de la SCM para el problema se muestran en el cuadro 2.30.
- Por otro lado se calcula la suma de cuadrados totales (SCT) con la ecuación (2.14).

$$SCT = \sum_{i=1}^{k}\sum_{j=1}^{b} \left(\bar{X}_{ij} - \bar{\bar{X}}\right)^2 \qquad (2.14)$$

Para $i = 1, 2, 3, \ldots k$ tratamientos.

$j = 1, 2, 3, \ldots b$ bloques.

$$SCT = (25 - 36.267)^2 + (21 - 36.267)^2 + (27 - 36.267)^2 + (35 - 36.267)^2 + (33 - 36.267)^2 + (31 - 36.267)^2 + ... + (45 - 36.267)^2 = 806.93$$

Cuadro 2.30 Resultados del cálculo de la SCM.

	Sistema de atención a clientes (tratamiento)		
Años de experiencia del cajero	Oracle (clientes por hora)	Total (clientes por hora)	Adabas (clientes por hora)
1	25	21	27
2	35	33	31
3	39	39	42
4	37	41	38
5	45	46	45
$\bar{x}_j$	36.2	36.0	36.6
$n_j(\bar{x} - \bar{\bar{x}})_2$	0.0222	0.3556	0.5556
	SCM = 0.933		

- La suma de cuadrados dentro de las muestras (SCDM) se calcula como:

  SCDM = SCT − SCM − SCB (2.15)

  SCDM = 806.93 − 0.9333 − 764.93 = 41.07
- En 2.15 hay b bloques y k tratamientos; es decir, hay $n = bk$ observaciones.
- Los grados de libertad para cada una de las sumas de cuadrados son:

  SCDM = SCT − SCM − SCB

  $(b - 1)(k - 1) = (n - 1) - (k - 1) - (b - 1)$
- Para el problema, los grados de libertad para cada una de las sumas de cuadrados son:

  $(5 - 1)(3 - 1) = (15 - 1) - (3 - 1) - (5 - 1)$

  $8 = (14) - (2) - (4)$
- El cuadrado medio entre muestras (varianza de la media entre muestras), lo mismo que el **Anova** de un factor son la suma de sus cuadrados dividido entre sus grados de libertad.

$$\sigma^2_{Ec} = \frac{SCM}{k-1} \qquad (2.15)$$

- Para el problema:

$$\sigma^2_{EC} = \frac{0.9333}{3-1} = 0.467$$

- En el análisis de varianza con dos factores, la varianza dentro de las muestras está definida como:

$$\sigma^2_{DC} = \frac{SCDM}{(b-1)(k-1)} \qquad (2.16)$$

- Para el problema:

$$\sigma^2_{DC} = \frac{41.07}{(5-1)(3-1)} = 5.133$$

- Y la varianza entre bloques (cuadrado medio del bloque) σ^2_b se calcula con la ecuación 2.17.

$$\sigma^2_b = \frac{SCB}{b-1} \qquad (2.17)$$

- Para el problema:

$$\sigma^2_b = \frac{764.93}{5-1} = 191.233$$

Prueba de hipótesis para Anova con dos factores mediante el *estadístico F*

Una vez que contamos con las tres estimaciones de la varianza de la población, el siguiente paso es compararlas mediante el cálculo de los cocientes siguientes:

$$F_1 = \frac{Varianza\ entre\ las\ medias\ muestrales\ (\sigma^2_{EC})}{Varianza\ dentro\ de\ las\ muestras\ (\sigma^2_{DC})} \quad (2.18)$$

$$F_2 = \frac{Varianza\ entre\ bloques\ (\sigma^2_{b})}{Varianza\ dentro\ de\ las\ muestras\ (\sigma^2_{DC})} \quad (2.19)$$

Estos cocientes calculan un estadístico *F* para probar la hipótesis de que las medias entre tratamientos son iguales (F_1), y que las medias entre los bloques son iguales (F_2).

Si las medias entre los bloques (filas) son iguales (el promedio de operaciones en caja es similar independientemente de la experiencia de los cajeros), entonces este factor no es determinante para explicar el comportamiento de la variable dependiente (promedio de operaciones en caja por hora) es decir, primero debemos realizar una prueba de hipótesis para filas. Por tanto, para nuestro problema, la hipótesis a probar es:

$$H_0: \mu_1 = \mu_2 = ... \mu_5$$

- La media de operaciones en cada caja (operaciones por hora) es similar sin importar los años de experiencia de los cajeros en sistemas.

$$H_1: \mu_1 \neq \mu_2 \neq ... \mu_5$$

- Las medias de operaciones en cada sistema son distintas y dependen de cuántos años de experiencia tenga el cajero en el manejo de sistemas.
- Obtenemos el valor de F_2 para realizar la prueba de hipótesis por bloques (filas).

$$F_{2=} \frac{\left(\sigma^2_b\right)}{\left(\sigma^2_{DC}\right)} = \frac{191.233}{5.133} = 37.255$$

- Si deseamos probar la hipótesis a un nivel de significancia del 5%, entonces el valor de F_2 deberá buscarse en tablas o bien calcularse con ayuda de Excel, con $(b-1)$ grados de libertad en el numerador y $(b-1)(k-1)$ grados de libertad en el denominador:

$$F_{2(0.05,\ 4,\ 8)} = 3.84$$

- Con:

Excel, =distr.f.inv(0.05, 4, 8)

- Obtenemos un valor de:

3.8379

- Nuestra regla de decisión en este caso sería:

NO rechazar la hipótesis nula (H_0), si $F_2 < 3.84$.

Rechazarla, si $F_2 > 3.84$.

- Debido a que F_2 calculada ($F_2 = 37.255$) es mayor que el valor en tabla (3.84) debemos rechazar la hipótesis nula y quedarnos con la hipótesis alternativa H_1; es decir, la media de operaciones en cada caja (operaciones por hora) realizadas en los diferentes sistemas son distintas y dependen de los años de experiencia que tenga el cajero en el manejo de sistemas.

- Debido a que el factor años de experiencia en sistemas es un elemento importante, entonces se procede a realizar un segundo análisis (**Two-Way Anova**) sobre los datos.
- Si el resultado de esta primera prueba nos hubiera indicado que debemos aceptar la hipótesis nula, entonces el analista deberá replantear su problema y realizar un análisis de varianza de un factor.
- Por tanto, para el problema de estudio, significaría que el factor experiencia del cajero en sistemas no es importante y que no es necesario crear los bloques de datos por años de experiencia.
- Ahora bien, como el factor bloqueado afecta los datos del problema, entonces el analista deberá realizar una segunda prueba de hipótesis.
- Por tanto, para nuestro problema, como la experiencia del cajero en sistemas sí afecta el número de operaciones por hora, se probará ahora una segunda hipótesis:

El número de operaciones que se pueden manejar por hora en cada sistema es igual siempre y cuando los cajeros que prueban cada sistema tengan los mismos años de experiencia.

- Es decir, el número promedio de operaciones en cada sistema es similar, por lo que el banco puede seleccionar cualquiera de ellos una vez hecha la consideración de los años de experiencia en sistemas que tiene el cajero.
- Este segundo análisis tiene como objetivo probar:

  H_0: $\mu_1 = \mu_2 = \mu_3$; es decir, la media de operaciones (operaciones por hora) es similar en cada sistema.

  H_1: $\mu_1 \neq \mu_2 \neq \mu_3$; es decir, las medias de operaciones son distintas en cada sistema.

- Para probar esta segunda hipótesis, obtenemos el valor de F_1 que permite realizar una prueba de hipótesis por tratamientos (columnas).

$$F_1 = \frac{Varianza\ entre\ las\ medias\ muestrales\ (\sigma^2_{EC})}{Varianza\ dentro\ de\ las\ muestras\ (\sigma^2_{DC})}$$

$$F_1 = \frac{(\sigma^2_{EC})}{(\sigma^2_{DC})} = \frac{0.467}{5.133} = 0.091$$

- Si deseamos probar esta segunda hipótesis a un nivel de significancia de 5%, entonces el valor de F_1 deberá buscarse en tablas o calcularse con ayuda de Excel con $(k - 1)$ grados de libertad en el numerador y $(b - 1)\ (k - 1)$ grados de libertad en el denominador:

$$F_{1(0.05,\ 2,\ 8)} = 4.46$$

- Con Excel, =DISTR.F.INV(0.05, 2, 8):

Obtenemos un valor de 4.459

- Nuestra regla de decisión en este segundo caso sería:

NO rechazar la hipótesis nula (H_0) si $F_1 < 4.46$.

Rechazarla si $F_1 > 4.46$.

- Debido a que F_1 calculada ($F_1 = 0.091$) es menor que el valor en tablas (4.46) debemos aceptar la hipótesis nula; es decir, la media de operaciones en cada sistema (operaciones por hora) son iguales siempre y cuando los cajeros cuenten con los mismos años de experiencia en la operación de sistemas.
- Con base en esto, el banco puede decidir comprar cualquiera de los sistemas que le han presentado sus proveedores.

Cuadro resumen del análisis de varianza con dos factores

Al igual que en el análisis de varianza para un factor, estos cálculos del ANOVA con dos factores y el formato del cuadro para análisis se pueden resumir en el cuadro 2.31.

Cuadro 2.31 Cálculos del ANOVA con dos factores.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	F	p
Entre muestras (tratamientos)	SCM	$k-1$	σ^2_{EC}	$\frac{(\sigma^2_{EC})}{(\sigma^2_{DC})}$	Significancia (sig.)
Entre bloques	SCB	$b-1$	σ^2_b	$\frac{(\sigma^2_b)}{(\sigma^2_{DC})}$	Significancia (sig.)
Dentro de muestras (error)	SCDM	$(b-1)(k-1)$	σ^2_{DC}		
Variación total	SCM + SCB + SCDM	$n-1$			

Nota: $n = n_1 + n_2 + \ldots n_k$, suma del número de datos de todas las muestras.
K = número de muestras o tratamientos. b = número de bloques.

- Para nuestro problema el cuadro resumen del análisis de varianza con dos factores se muestra en el cuadro 2.32.

Cuadro 2.32 Resumen del análisis de varianza con dos factores.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	F	P
Entre muestras (tratamientos)	0.933	2	0.467	0.091	0.914
Entre bloques	764.93	4	191.233	37.255	0.000
Dentro de muestras (error)	41.07	8	5.133		
Variación total	806.93	14			

En el cuadro 2.33, resultado del proceso con SPSS, se muestra el resumen del análisis de varianza con dos factores y el cálculo del estadístico p para ambas pruebas de hipótesis.

Cuadro 2.33 Test of between-subjects effects.

Dependent variable: Número de operaciones bancarias por hora

Source		Type III Sum of squares	df	Mean Square	F	Sig.
Sistema	Hypothesis	.933	2	0.467	.091	.914
	Error	41.067	8	5.133[a]		
Experiencia	Hypothesis	764.933	4	191.233	37.253	.000
	Error	41.067	8	5.133[a]		
Sistema * Experiencia	Hypothesis	41.067	8	5.133	.	.
	Error	.000	0	[b]		

a. MS(Sistema * Experiencia)
b. MS(Error)

Para la primera hipótesis del problema:

- El valor p es menor que alfa ($0.000 < 0.05$).
- Por lo que debe rechazarse la hipótesis nula.
- La experiencia sí tiene significancia en el manejo de los sistemas.
- Con esto se corrobora el resultado encontrado con la prueba F_2.

Por otro lado, para la segunda hipótesis del problema:

- El valor p es mayor que alfa ($0.914 > 0.05$).
- Por lo que debe aceptarse la hipótesis nula.
- Es decir, el promedio de operaciones por hora es similar en cualquiera de los sistemas si se consideran cajeros con la misma experiencia (*véase* el cuadro 2.32).

Ejemplo 2.6

La Dirección de Recursos Humanos (RH) de BIMBO está probando un nuevo sistema de evaluación para sus gerencias en una planta de la Ciudad de México, en el que se considera una escala de calificación de 10 a 50 puntos. Se seleccionan aleatoriamente cinco empleados de la planta y se les pide que evalúen la actuación de los gerentes de producción, finanzas, mercadotecnia y almacenes (*véase* el cuadro 2.34).

Cuadro 2.34 **Evaluación para las gerencias en la planta BIMBO.**

	Gerente (tratamiento)			
Empleado	Producción	Finanzas	Mercadotecnia	Almacenes
1	38	35	31	46
2	36	32	29	45
3	20	17	13	37
4	39	38	28	50
5	20	20	14	40

- La dirección de RH desea saber si existen diferencias en las calificaciones promedio de los cuatro gerentes.
- Sospecha que el sistema permite a todos los empleados evaluar igual a todos los gerentes por lo que decide realizar un análisis de varianza con dos factores.

Solución

- Calculamos primero la suma de cuadrados de bloques (SCB) como se muestra en el cuadro 2.35.

Cuadro 2.35 **Suma de cuadrados de bloques (SCB)**

	Gerente (tratamiento)					
Empleado	Producción	Finanzas	Mercadotecnia	Almacenes	$\bar{x}i$	$k(\bar{x}-\bar{\bar{x}})^2$
1	38	35	31	46	37.50	148.84
2	36	32	29	45	35.50	67.24
3	20	17	13	37	21.50	372.49
4	39	38	28	50	38.75	216.64
5	20	20	14	40	23.50	249.64
			$\bar{\bar{x}}$	31.4	SCB =	1054.30

- Calculamos la suma de cuadrados entre muestras (SCM) con el numerador de la ecuación 2.5 (*véase* el cuadro 2.36).

Cuadro 2.36 Suma de cuadrados entre muestras (SCM).

	Gerente (tratamiento)			
Empleado	Producción	Finanzas	Mercadotecnia	Almacenes
1	38	35	31	46
2	36	32	29	45
3	20	17	13	37
4	39	38	28	50
5	20	20	14	40
$\bar{x}_j$	30.6	28.4	23	43.6
$n_j\left(\bar{x}_j - \bar{\bar{x}}\right)^2$	3.2	45	352.8	744.2
		SCM =	1145.2	

- Se calcula la suma de cuadrados totales (SCT) con la ecuación (2.14).

$$\text{SCT} = (38 - 31.4)^2 + (35 - 31.4)^2 + (31 - 31.4)^2 + ... + (40 - 31.4)^2 = 2284.8$$

- Se calcula la suma de cuadrados dentro de las muestras (SCDM) como:

$$\text{SCDM} = \text{SCT} - \text{SCM} - \text{SCB}$$

$$\text{SCDM} = 2284.8 - 1145.2 - 1054.30 = 85.3$$

- El resumen del análisis de dos factores para este problema queda definido en el cuadro 2.37.

Cuadro 2.37 Resumen del análisis de dos factores.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	F	p
Entre muestras tratamientos	1145.2	3	381.73	53.70	0.000
Entre bloques	1054.3	4	263.58	37.08	0.000
Dentro de muestras (error)	85.3	12	7.11		
Variación total	2284.8	19			

- En principio el director desea saber si hay una diferencia significativa entre las calificaciones promedio que han dado cada uno de los cinco empleados (filas).
- Las hipótesis en este caso son:

$$H_0: \mu_1 = \mu_2 = ... \mu_5$$

- La media de calificación de cada empleado es similar.

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 ... \mu_5$$

- Las medias de calificación de cada empleado son distintas.
- El analista desea utilizar un nivel de significancia de 5%, entonces el valor F_2 buscado en tablas o calculado con EXCEL es:

$$F_{2(0.05,\ 4,\ 12)} = 3.26$$

- Con EXCEL:

=DISTR.F.INV(0.05, 4, 12), obtenemos un valor de 3.2592

- Nuestra regla de decisión en este caso sería:

NO rechazar la hipótesis nula (H_0) si $F_2 < 3.26$. Rechazarla si $F_2 > 3.26$.

- Debido a que F_2 calculada ($F_2 = 37.08$) es mayor que el valor en los cuadros (3.26) debemos rechazar la hipótesis nula y quedarnos con la hipótesis alternativa H_1, es decir, la media de calificación de cada empleado es distinta y requerimos de un bloqueo del factor.
- Ahora bien, como el factor bloqueado afecta los datos del problema, entonces el analista debe proceder a realizar una segunda prueba de hipótesis (**Two-Way Anova**).
- El director puede probar la hipótesis inicial respecto a que las calificaciones promedio recibidas por las gerencias son similares.

$$H_0: \mu_1 = \mu_2 = \ldots \mu_4$$

- La media de calificación de cada gerencia es similar en este sistema de evaluación.

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \ldots \mu_4$$

- Las medias de calificación de cada gerencia son distintas en este sistema de evaluación.
- El analista debe buscar a un nivel de significancia de 5%, el valor F_1 en tablas o calcularlo con Excel.

$$F_{1(0.05,\,3,\,12)} = 3.49$$

- Con Excel:

=DISTR.F.INV(0.05, 3, 12), obtenemos un valor de 3.4903

- Nuestra regla de decisión en este segundo caso sería:

NO rechazar la hipótesis nula (H_0) si $F_1 < 3.49$. Rechazarla si $F_1 > 3.49$.

- Debido a que F_1 calculada ($F_1 = 53.70$) es mayor que el valor en tablas (4.49) debemos rechazar la hipótesis nula; es decir, las medias de calificación de cada gerencia son distintas en este sistema de evaluación.

Análisis de varianza con dos factores mediante SPSS

Un análisis de varianza con dos factores (**Two-Way Anova**) mediante el uso del paquete Statistical Package for Social Sciences (SPSS) versión 16.0 puede elaborarse realizando los pasos siguientes:

1. Primero se definen las variables de análisis, al retomar el ejemplo anterior, las variables de análisis quedarían definidas en la ventana de vista de variables (**variable view**) como:

   Empleado (I, II, III, IV y V), Gerencia (1, Producción; 2, Finanzas; 3, Mercadotecnia y 4, Almacenes) y Calificación del empleado a cada gerencia (*véase* la figura 2.6 del editor de SPSS).

anova2fej.sav [DataSet2] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Dec...	Label	Values	Missing	Columns	Align	Measure
1	Empleado	Numeric	1	0	Empleado	{1, I}...	None	8	Center	Scale
2	Gerencia	Numeric	1	0	Gerencia a evaluar.	{1, Producción...	None	8	Center	Scale
3	Calificación	Numeric	2	0	Calificación	None	None	8	Center	Scale
4										
5										
6										
7										
8										

Figura 2.6 Variables de análisis.

2. Se capturan los datos en la ventana de vista de datos (**Data View**), como lo muestra la figura 2.7.
3. Se solicita el cálculo del análisis de varianza con dos factores, para ello, del menú **Analize**, submenú, **General Linear Model**, seleccionamos la rutina Univariate… (*véase* la figura 2.8).

anova2fej.sav [DataSet2] - SPSS Data Editor

	Empleado	Gerencia	Calificación	var	var	var
1	1	1	38			
2	1	2	35			
3	1	3	31			
4	1	4	46			
5	2	1	36			
6	2	2	32			
7	2	3	29			
8	2	4	45			
9	3	1	20			
10	3	2	17			
11	3	3	13			
12	3	4	37			
13	4	1	39			
14	4	2	38			
15	4	3	28			
16	4	4	50			
17	5	1	20			
18	5	2	20			
19	5	3	14			
20	5	4	40			
21						

Figura 2.7 Captura de los datos en la ventana de vista de datos.

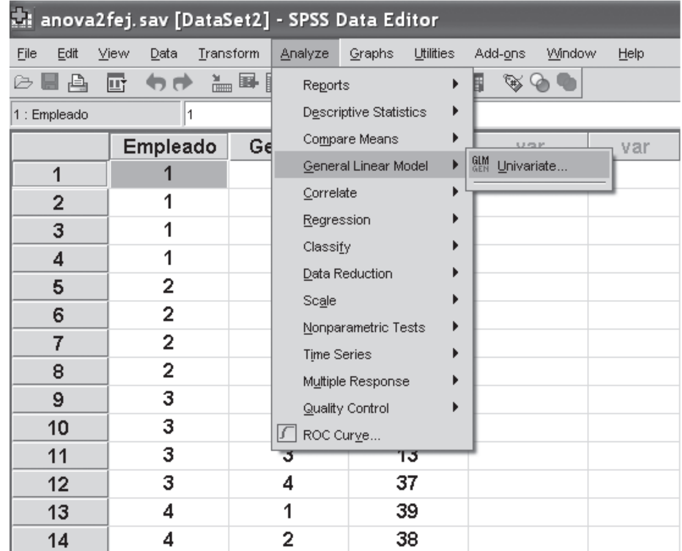


Figura 2.8 Cálculo del análisis de varianza con dos factores.

4. Seleccionamos la variable dependiente (**Calificación de la gerencia**), la variable independiente o factor fijo (**Gerencia**) y el segundo factor o factor aleatorio (**Empleado**) como se muestra en la figura 2.9. En la sección de Modelo (**Model**…) eliminar el cálculo de la intersección (**Include intercept in model**).

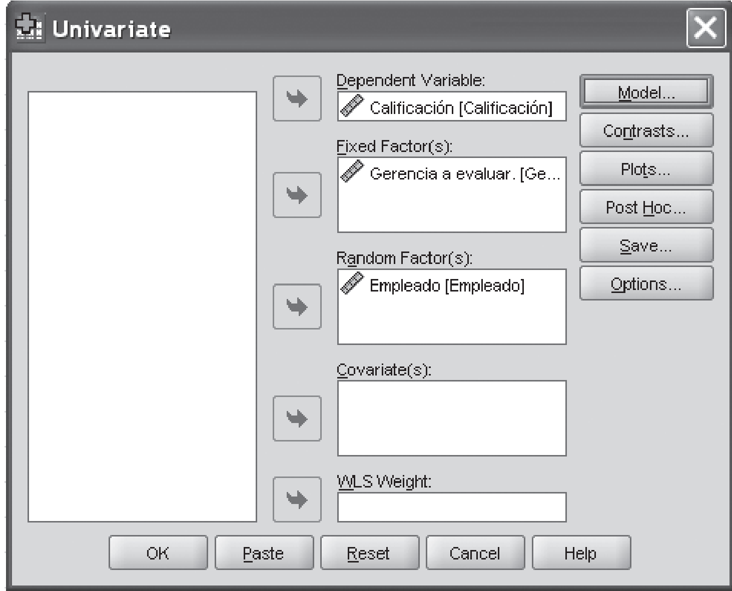


Figura 2.9 Se selecciona la variable dependiente, la independiente o factor fijo y el segundo factor o factor aleatorio.

5. Solicitar el cálculo mediante **OK** en la ventana de **Univariate** para obtener los resultados del análisis de varianza con dos factores (*véase* el cuadro 2.38).

Cuadro 2.38 Se solicita el cálculo mediante *OK* en la ventana de *Univariate*.

Between Subjets Factors

		Value label	*N*
Gerencia a evaluar	1	Producción	5
	2	Finanzas	5
	3	Mercadotecnia	5
	4	Almacenes	5
Empleado	1	I	4
	2	II	4
	3	III	4
	4	IV	4
	5	V	4

Test of between subjets effects

Dependent variable: calificación

Source		*Type III Sum or squares*	*df*	*Mean square*	*F*	*Sig.*
Gerencia	Hypothesis	1145.200	3	381.733	53.702	0.000
	Error	85.300	12	7.108[a]		
Empleado	Hypothesis	1054.300	4	263.575	37.080	0.000
	Error	85.300	12	7.108[a]		
Gerencia * Empleado	Hypothesis	85.300	12	7.108		
	Error	0.000	0	[b]		

a. MS(Gerencia * Empleado).
b. MS(Error).

Análisis de factores

El análisis de factores es un procedimiento estadístico fundamentado en el análisis de varianza, por lo que esta herramienta estadística permite probar simultáneamente el efecto de dos factores; es decir, evaluar dos factores de interés al mismo tiempo.

En el análisis de varianza con dos factores, uno es bloqueado durante el procedimiento de cálculo para eliminar su impacto, pero en el análisis factorial, como ya indicamos, ambos factores son evaluados al mismo tiempo, ninguno es bloqueado ya que nos interesa medir el impacto de ambos.

En el análisis factorial se debe efectuar una prueba de efectos principales por cada factor. Esta prueba se aplica sobre ambos factores para determinar si los niveles diferentes del factor influyen en las unidades de manera diversa. Si no se encuentran efectos principales para un factor, la hipótesis nula no debe rechazarse.

En este tipo de análisis, cada factor tiene más de un nivel. Si A y B son los factores de interés, entonces A tiene "*a*" niveles y B, "*b*" niveles.

Por ejemplo, si el factor A tiene 3 niveles (I, II y III) y el B 3 (1, 2 y 3), entonces se tiene un diseño factorial 3x3. Cada combinación (celda) se denomina como un tratamiento (*véase* el cuadro 2.39).

Cuadro 2.39 Cada combinación (celda) se denomina como un tratamiento.

		Factor B		
		1	2	3
Factor A	I	dato$_{I11}$ dato$_{I12}$ dato$_{I13}$	dato$_{I21}$ dato$_{I22}$ dato$_{I23}$	dato$_{I31}$ dato$_{I32}$ dato$_{I33}$
	II	dato$_{II11}$ dato$_{II12}$ dato$_{II13}$	dato$_{II21}$ dato$_{II22}$ dato$_{II23}$	dato$_{II31}$ dato$_{II32}$ dato$_{II33}$
	III	dato$_{III11}$ dato$_{III12}$ dato$_{III13}$	dato$_{III21}$ dato$_{III22}$ dato$_{III23}$	dato$_{III31}$ dato$_{III32}$ dato$_{III33}$

i = I, II y III para el factor A

j = 1, 2 y 3 para el factor B

k = número de observación dentro de la celda, k = 1,2,3, …, r

- Para realizar el análisis factorial, debe aparecer más de una observación (dato) en cada celda.
- El número de observaciones dentro de una celda se denomina número de replicaciones “r”.
- En el método que se describe en este libro cada celda debe tener el mismo número de observaciones.
- Un número desigual va más allá del alcance de este texto.

En este método se puede identificar toda interacción que pueda presentarse entre los dos factores y dicha interacción sería imposible detectarla si los experimentos se realizaran por separado o si cada celda tuviera sólo una observación.

Se dice que la interacción existe si un nivel del factor *A* funciona de manera diferente (mejor o peor) con niveles diferentes del factor *B*. Esta interacción puede detectarse analizando las diferencias promedio entre los niveles de un factor en relación con los niveles diferentes del otro factor.

Si estas diferencias promedio son las mismas en todos los niveles de ambos factores, no existe interacción y, se dice que los efectos de estos factores son aditivos.

Un método más preciso para detectar la interacción consiste en aplicar prueba de hipótesis y se pueden identificar tres hipótesis que deben probarse:

H_0: Las medias de las filas son iguales (prueba de efectos principales del factor 1 (A)).

H_0: Las medias de las columnas son iguales (prueba de efectos principales del factor 2 (B)).

H_0: No hay interacción presente.

Las hipótesis alternativas (H_1) para cada prueba se plantean al contrario:

- Al igual que el análisis de varianza para un factor y con dos factores, este método permite desglosar las sumas de cuadrados y construir la tabla factorial para probar las hipótesis.
- El análisis factorial tiene la ventaja de ser menos costoso.
- Se pueden estudiar dos factores en un solo experimento en lugar de realizar dos pruebas independientes.

Ejemplo 2.7

El programa del doctorado en Ciencias Sociales de la UAM-X realiza un examen de aptitudes de 70 puntos (el mínimo aceptable es de 25) a candidatos que tienen licenciatura en Sociología, Ingeniería y Comunicación Social, para determinar los candidatos idóneos.

Para presentar este examen, la universidad imparte tres programas de capacitación:

> Revisión de conocimientos en 20 horas, un seminario de una semana (40 horas) y un curso propedéutico de 11 semanas.

Se selecciona al azar seis estudiantes de cada una de las tres licenciaturas, dos de cada uno de los tres programas de capacitación que impartió la universidad y cuyas calificaciones obtenidas por los candidatos se muestran en el cuadro 2.40, por lo que con base en éstas se desea saber:

- ¿Difieren los programas de capacitación en cuanto a sus efectos sobre las calificaciones de la evaluación de aptitudes?
- ¿Difiere la formación académica (licenciatura) en cuanto a sus efectos sobre las calificaciones de la evaluación de aptitudes?
- ¿Se desempeñan mejor los egresados de alguna licenciatura en determinado programa de capacitación, y los de otra licenciatura en otro tipo de programa de capacitación?

Cuadro 2.40 **Cada combinación (celda) se denomina como un tratamiento.**

		Factor: Licenciatura		
		1. Sociología	2. Ingeniería	3. Comunicación Soc.
Factor Programa de Capacitación	I	58	46	40
	20 hrs.	50	54	48
	II	46	56	42
	40 hrs.	54	62	48
	III	56	60	48
	11 semanas	60	58	41

Procedimiento de cálculo del análisis de factores

El procedimiento de cálculo del análisis de factores es similar a los dos procedimientos anteriores ya que nuevamente se divide la suma de cuadrados y los grados de libertad que corresponde a cada fuente. La fórmula para dividir la suma de cuadrados en este método cuando se manejan dos factores es:

$$\text{SCT} = \text{SCA} + \text{SCB} + \text{SCAB} + \text{SCE} \qquad (2.20)$$

En donde la partición de la suma de cuadrados y de los grados de libertad se define con:

a = Número de niveles del factor A.

b = número de niveles del factor B.

r = Número de réplicas (observaciones en cada bloque).

n_T = Número total de observaciones (datos) del experimento.

X_{ijk} = Observación de la k-ésima réplica tomada del tratamiento i del factor A y del tratamiento j del factor B.

$\overline{X}_i$ = Promedio de la muestra de las observaciones en el tratamiento i (factor A).

$\overline{X}_j$ = Promedio de la muestra de las observaciones en el tratamiento j (factor B).

$\overline{\overline{X}}$ = Gran media de las muestras para n_T observaciones.

- Suma de cuadrados del total (SCT):

$$SCT = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} \left(x_{ijk} - \bar{\bar{x}} \right)^2 \qquad (2.21)$$

$$SCT = (58 - 51.5)^2 + (50 - 51.5)^2 + (46 - 51.5)^2 + ... + (41 - 51.5)^2 = 824.50$$

- Suma de cuadrados para el factor A (SCA):

$$SCA = br \sum_{i=1}^{a} \left(\bar{X}_1 - \bar{\bar{X}} \right)^2 \qquad (2.22)$$

$$SCA = [(3)(2)[(49.333\text{-}51.5)]^2 + (51.333 - 51.5)^2 - (53.833 - 51.5)^2] = 61.0$$

- Suma de cuadrados para el factor B (SCB):

$$SCB = ar \sum_{j=1}^{b} \left(\bar{X}_1 - \bar{\bar{X}} \right)^2 \qquad (2.23)$$

$$SCB = [(3)(2)[(54 - 51.5)]^2 + (56 - 51.5)^2 + (44.5 - 51.5)^2] = 453.0$$

- Suma de cuadrados para la interacción (SCAB):

$$SCT = r \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{\bar{X}} \right)^2 \qquad (2.24)$$

$$SCAB = [(2)[(54 - 49.333 - 54 + 51.5)]^2 + (50 - 49.333 - 56 - 51.5)^2 + ... + (44.5 - 53.833 - 44.5 - 51.2)^2] = 112.0$$

- Suma de cuadrados debido al error (SCE):

  SCE = SCT − SCA − SCB − SCAB

  SCE = 824.5 − 61.0 − 453.0 − 112.0 = 198.5

En el cuadro 2.41 elaborado con una hoja electrónica de EXCEL se muestra un resumen de los cálculos anteriores:

Cuadro 2.41 Resumen de cálculos.

		Factor: Licenciatura				Gran Media
		1. Sociología	2. Ingeniería	3. Comunicación Soc.		51.5
	I	58	46	40	media de	
	20 hrs.	50	54	48	medias filas	SCT
Factor	media	54	50	44	49.333	824.50
Programa	II	46	56	42		
de	40 hrs.	54	62	48		SCA
Capacitación	media	50	59	45	51.333	61.00
	III	56	60	48		
	11 semanas	60	58	41		SCB
	media	58	59	44,5	53.833	453
	media					SCAB
	de medias	54	56	44,5		112
	columnas					

Los cálculos se pueden mostrar en el cuadro 2.42:

Cuadro 2.42 Resumen de análisis de factores.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	*F*	*p*
Factor A	SCA	$a - 1$	VFA=SCA/ $(a - 1)$	VFA/VE	Significancia (sig.)
Factor B	SCB	$b - 1$	VFB=SCB/$(b - 1)$	VFB/VE	Significancia (sig.)
Interacción	SCAB	$(a - 1)(b - 1)$	VFAB=SCAB/$(a - 1)(b - 1)$	VFAB/VE	Significancia (sig.)
Error	SCE	$ab\,(r - 1)$	VE=SCE/ab $(r - 1)$		
Variación total	SCA + SCB + SCAB + SCE	$n_T - 1$			

- Para el ejemplo, el cuadro resumen del análisis de factores se muestra en el cuadro 2.43 y una salida con el programa SPSS de este problema en el 2.44.

Cuadro 2.43 Análisis de factores.

Fuente	Suma de cuadrados	Grados de libertad	Varianza de la media	*F*	*p*
Capacitación	61.0	2	30.5	1.383	0.299
Licenciatura	453.0	2	226.5	10.270	0.005
Interacción	112.0	4	28.00	1.270	0.350
Error	198.5	9	22.06		
Variación total	824.5	17			

Cuadro 2.44 Salida con el programa SPSS.

Test of between subjets effects					
Dependent variable: calificación					
Source	*Type III sum or squares*	*df*	*Mean square*	*F*	*Sig.*
Model	48366.500[a]	9	5374.056	243.660	0.000
Programa	61.000	2	30.500	1.383	0.299
Licenciatura	453.000	2	226.500	10.270	0.005
Programa* Licenciatura	112.000	4	28.000	1.270	0.350
Error	198.500	9	22.056		
Total	48565.000	18			

[a.] R squared = 0.996 (adjusted R squared = 0.992)

Con base en la información de los cuadros 2.43 y 2.44 podremos responder a nuestras preguntas de investigación.

- ¿Difieren los programas de capacitación en cuanto a sus efectos sobre las calificaciones de la evaluación de aptitudes?:
  - **El valor crítico de *F* con $\alpha = 0.05$, con dos grados de libertad en el numerador y nueve en el denominador es 4.26.**

- Puesto que la F (1.383) para los programas de capacitación es menor que el valor crítico no podemos rechazar la hipótesis nula.
- El valor $p = 0.299$ nos confirma lo anterior y por lo tanto en este punto debemos concluir que no hay diferencias importantes entre los tres programas de capacitación para el examen de aptitudes del doctorado.

- ¿Difiere la formación académica (licenciatura) en cuanto a sus efectos sobre las calificaciones de la evaluación de aptitudes?:
  - En este caso el valor calculado de $F = 10.27$ es mayor al valor en las tablas (4.26), lo que nos indica que sí hay diferencias significativas entre la formación académica que afecta las calificaciones del examen.
  - El valor $p = 0.005$, confirma la afirmación anterior ya que $0.005 < 0.05$, lo que nos indica que debemos rechazar la hipótesis nula de este factor.

- ¿Se desempeñan mejor los egresados de alguna licenciatura en determinado programa de capacitación, y los de otra licenciatura en otro tipo de programa de capacitación?:
  - El valor en los cuadros de la F con $\alpha = 0.05$, con cuatro grados de libertad en el numerador y nueve en el denominador es 3.63.
  - Puesto que la F (1.270) para la interacción (programas-licenciaturas) es menor que el valor crítico no tenemos motivos para creer que los tres programas de capacitación que se imparten son distintos en cuanto a preparar a los estudiantes de las diferentes licenciaturas que desean ingresar a este programa.

Problemas

1. El departamento de producción económica de la UAM-X está comparando los salarios iniciales de sus egresados en Administración, los cuales dependen del área de especialización seleccionada (Finanzas, Mercadotecnia, Recursos Humanos y Producción). Se toma una muestra de 28 egresados. A un nivel de 5%, determine si hay diferencia en los salarios promedio de los egresados de esta licenciatura (*véase* el cuadro 2.45).

Cuadro 2.45 Comparación de salarios iniciales de egresados de la UAM-X.

	Especialización			
Egresados	Finanzas	Mercadotecnia	Recursos Humanos	Producción
1	23.2	23.3	22.1	22.2
2	24.7	22.1	19.2	22.1
3	24.2	23.4	21.3	23.2
4	22.9	24.2	19.8	21.7
5	25.2	23.1	17.2	20.2
6	23.7	22.7	18.3	22.7
7	24.2	22.8	17.2	21.8

Nota: Los ingresos están dados en miles de pesos mensuales.

2. Farmacia de Similares desea comparar la producción diaria promedio de suministros médicos de sus tres plantas de la zona central del país. Se recolectaron los datos de los últimos nueve días en cada planta. A un nivel de 10% existen diferencias en las medias de producción de sus plantas (*véase* el cuadro 2.46).

Cuadro 2.46 Producción diaria promedio de suministros médicos.

	Planta		
Muestra	Toluca	Puebla	Cuernavaca
1	10	15	12
2	12	17	17
3	15	18	15
4	18	12	15
5	9	13	18
6	17	11	12
7	15	12	13
8	12	11	14
9	18	12	14

Nota: Producción en miles de unidades.

3. La casa de bolsa MONEX México desea determinar si el promedio de comisiones diarias que ganan sus corredores cambia según el día de la semana. Se cuenta con los datos de siete de sus corredores que laboraron la semana pasada, por tanto determine (*véase* el cuadro 2.47).
   a) Si hay cambio en el promedio de comisiones por día de la semana,
   b) ¿Qué días parecen ganar más?
   c) Utilice el criterio de Tukey.

Cuadro 2.47 Promedio de comisiones diarias que ganan sus corredores cambia según el día de la semana.

Corredor	Lunes	Martes	Miércoles	Jueves	Viernes
1	2100	2800	1100	1500	2500
2	2600	2100	1400	1400	2300
3	2400	1900	1200	1200	2600
4	3200	1500	1000	1200	2800
5	2500	1200	1000	1600	2400
6	2600	1000	1200	1300	2500
7	2400	1300	1500	1800	2900

Nota: Comisiones en pesos.

4. Se registran las ventas diarias de tres tiendas WALL-MARK en la zona sur de la ciudad de México durante cinco días. El gerente regional desea saber si las ventas promedio en estas tiendas son similares, por tanto determine (*véase* el cuadro 2.48):
   a) A un nivel de significancia de 5% a que conclusión llega.
   b) ¿Qué tienda parecen ganar más?
   c) Utilice el criterio de Tukey.

Cuadro 2.48 Ventas diarias de tres tiendas WALL-MARK durante cinco días.

	Tienda		
Día	Miramontes	Coapa	Tlalpan
1	32	44	33
2	20	43	36
3	30	44	35
4	26	46	36
5	32	48	40

Nota: Miles de pesos.